DOCUMENT RESUME

ED 228 270 TM 830 118

AUTHOR Quellmalz, Edys S.
TITLE Designing Writing Assessments: Balancing Fairness, Utility, and Cost.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO CSE-R-188
PUB DATE 82
NOTE 24p.
PUB TYPE Information Analyses (070) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cost Effectiveness; Educational Assessment; *Evaluation Criteria; Evaluation Methods; Holistic Evaluation; *Minimum Competencies; *Minimum Competency Testing; Rating Scales; Scoring; Secondary Education; *Test Construction; *Writing Evaluation; *Writing Skills
IDENTIFIERS Fairness

ABSTRACT

Mounting concern for student achievement in writing has refocused attention on the features of writing assessment necessary to represent a student's skill fairly, usefully and economically. If writing tests are to fulfill their intended function, the writing assignments and evaluative criteria of large scale tests and instruction must interrelate. Current practices are increasingly criticized regarding relevance to realistic writing situations, utility for forming decisions about individual competence or program effectiveness, fairness, and legality for sanctioning exit requirements. State and district writing assessments should re-evaluate their methods considering alternatives proposed by recent writing theory and research. Specifying writing goals which distinguish between minimum functional goals and desirable goals of competence may improve the logic, utility and fairness of test procedures. Appropriate writing tasks should be designed to provide a full rhetorical context, and time to engage in all parts of the writing process. An integrated instructional system which targets particular writing elements as important basic competencies would involve teachers and evaluators in specification of rating criteria--whether holistic judgments or several separate analytic scores. The technical quality of rating criteria is a problem of scale stability and validity. Cost concerns should not outweigh concerns for fairness and utility. (CM)

DESIGNING WRITING ASSESSMENTS: BALANCING FAIRNESS, UTILITY, AND COST

Edys S. Quellmalz

CSE Report No. 188
1982

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The research reported herein was supported in whole or in part by a grant to the Center for the Study of Evaluation from the National Institute of Education, U. S. Department of Education. However, the opinions and findings expressed here do not necessarily reflect the position or policy of NIE and no official endorsement should be inferred.

TABLE OF CONTENTS

INTRODUCTION

To attain the fundamental goal of language competence, educators, students, and parents must have information describing the status and progress of language skills development. Mounting concern for student achievement in writing, one of the principal arenas of language development, has refocused the attention of policy makers, evaluators, instructors, and researchers on the features of writing assessment necessary to represent a student's writing skill fairly, usefully, and economically. While the relationship between procedures employed to evaluate writing in large scale testing and those used in the classroom has historically been tenuous, the requirements of minimum competency testing programs have stimulated research on methods to tighten the connection. These competency testing programs require school systems to assess the status of students' basic skill achievement, and then either to certify that minimal competencies have been attained or signal the need for remediation and provide repeated opportunities for students to pass comparable test forms. If these writing competency tests are to fulfil their intended function, then the writing assignments and evaluative criteria of large scale tests and classroom instruction must interrelate.

At present, many large scale writing tests bear little resemblance to students' classroom writing experiences. Many states and districts rely on multiple choice tests that measure sentence-level editing skills or passage comprehension. When writing samples are collected, the structure and topic of the writing assignment may call for information and strategies that vary considerably from students' experiences in and out of the classroom. Furthermore, writing samples are often scored rapidly and holistically by raters trained to varying levels of precision and accuracy. Students receive a single score purportedly representing the level of their writing competence.

Reactions of practitioners and researchers to such current practices are increasingly critical. They find many faults in current writing tests -- their logical and psychological relevance to realistic writing situations, their utility for informing decisions about individual competence or program effectiveness, their fairness to students and instruction, their legality for sanctioning exit requirements. This paper suggests that state and district writing assessments should re-evaluate their current methods for assessing student writing competence in light of these criticisms. An accumulating body of literature indicts many of the methods assessments now use that have been derived from custom, folklore, and adaptations of norm-referenced testing methodology that are inappropriate for the purposes of competency assessment. By examining the criticisms leveled at writing tests and considering alternatives proposed by recent writing theory and research, we may find solutions that will improve the fairness and utility of writing assessments, yet remain within reasonable economic bounds.

PROBLEM 1: SPECIFYING WRITING GOALS

Just what is "good" writing? For schools, a major conflict has been to distinguish between realistic characteristics of minimum competence, reasonable high school writing exit competence, and the competence of professional writers and "experts." A significant component in this controversy over "standards" has been the function various types of writing can and/or should have for the student. Thus the discourse aim or writing purpose of transactional writing has been identified by many school systems as functionally most relevant to the majority of students. At the lower grades, expressive writing has been viewed by some as valuable in its own right and by others as an educational vehicle for motivating writing that will increase fluency and sentence-level competence.

Clearly, the schools' definition of the target constrains the specific criteria that will provide logical and empirical evidence that the target has been hit. Currently, goals may relate to two competency levels, a *minimum* competency level targeted by most state and district minimum competency testing programs and a reasonably desirable *high school exit* competency level implied in many systems' curricular goals. Most competency programs emphasize transactional writing in the factual narrative, expository, or persuasive modes. Minimum program goals are often that students write a clear, coherent paragraph that makes a point and that exhibits few or no mechanical, sentence-level errors. For high school exit goals, English departments set their sights at the multi-paragraph, essay level, seeking writing that has a theme or point, that is coherent between, as well as within paragraphs, and that exhibits few sentence-level errors. While minimum goals generally specify functional writing, high school exit goals may expand the types of writing aims or purposes in which it *is desired* that students be competent. By distinguishing between minimum and desirable goals, school systems may be in a better position to defend the logic, utility, and fairness of focused test procedures.

PROBLEM 2: DESIGNING APPROPRIATE WRITING TASKS

Perhaps the most common controversy in the design of writing tests involves the relative merits of direct and indirect tasks. Indirect, usually multiple choice, measures have been defended by test publishers because of their economy and high correlations with essay scores (Godshalk, Swineford & Coffman, 1966; Breland & Braucher, 1977). Critics of multiple choice tests reject them on logical and psychological grounds. They argue that multiple choice tests present primarily editing tasks or comprehension tasks and t'at they therefore do not tap the same kinds of mental processes required by production tasks (Bourne, 1966; Quellmalz, 1978; Cooper, 1979). Recent empirical studies of students' scores on direct and indirect measures indicate considerably lower correlations between writing skill component scores derived from multiple choice and writing samples (Quellmalz & Capell, 1979; Quellmalz, Smith, Winters & Baker, 1980; Moss, Cole & Khampaliket, in press). Furthermore, Quellmalz and Capell found multiple choice test scores provided less distinctive information about underlying writing skill constructs or traits than did essay ratings (Quellmalz & Capell, 1979). In combination, these studies support contentions that direct and indirect measures tap different psychological processes. These data would also, of course, suggest that multiple choice test scores would not serve as fair or useful proxies for actual writing skill. At best, multiple choice tests seem to over-estimate skills (NAEP, 1981) since they measure skills presumably enroute to production skills (Skinner, 1957).

In addition to debate over the form of response required by writing tests, there is considerable disagreement about the appropriate structure of assignments used to prompt writing. Criticisms of writing tasks are that they do not present full rhetorical contexts that sufficiently inform students about the writing purpose, topic, audience, writer's role, and intended criteria (Britton, 1978; Cazden, 1974; Scribner & Cole, 1978; Florio, 1979). Research shows that writers' performance differs when writing in different discourse modes, e.g., exposition and narration (Veal & Tillman, 1971; Crowhurst, 1980; Quellmalz & Capell, 1979; Praeter & Padia, 1980; Baker & Quellmalz, 1980). Research also reveals that accessibility of information about an assigned topic affects the quality of students' writing (Baker & Quellmalz, 1980). Polin (1980) has found that when writers are given extended time and cues about the rhetorical demands of the task during planning or revision, some of them improve in various features of their work. In sum, studies of features of the writing task that influence students' writing performance suggest that variations within features such as mode of discourse (writing aim) topic, audience, time, and structural cues do present different psychological demands and therefore should be distinctly specified. To be clear and fair, the writing task should provide a full rhetorical context and time to engage in all parts of the writing process. The cost of developing well formed writing prompts is not high, particularly in comparison to the cost of erroneous inferences about competence made from assessments of writing that students generate in response to incomplete or ambiguous prompts.

PROBLEM 3: SPECIFYING SCORING CRITERIA AND TYPE OF RATING SCALE

Criteria employed for evaluating student writing vary along a number of dimensions: from qualitative to quantitative; from generai to specific; from comprehensive, full discourse features to isolated features; from vague guidelines to replicable, objective guidelines.

At the most qualitative, vague end of the continua are general impression scoring schemes where readers apply their own criteria co give the writing a single global score. Follman and Anderson's "Everyman" procedures (1967) and teachers' A-F grading schemes fall in this category. Still providing a single score or quality rating, but guided by slightly more descriptive and acknowledged criteria, are holistic rating schemes such as the ETS four or six-point scales which rank papers within a set. Teachers' use of a letter grade with some supporting comments might relate to this evaluation scheme. Some rating schemes are specific to discourse mode; others, like the primary trait rating method, are specific to discourse mode and the particular topic (Lloyd-Jones, 1977). The most detailed scales are analytic rating schemes referencing component features of the written product.

Where do these criteria come from? Criteria for these scales may be inferred from features commonly referenced by knowledgeable readers, they may be arbitrary, or theoretically- or empirically-based dimensions deemed important by the group designing the scheme. Analytic scales vary in the degree to which they comprehensively reference rhetorical, structural, and syntactic features, as well as the degree to which criteria for features are qualitative or quantitative. In an attempt to be comprehensive, the subscales of the Diederich Expository Scale range from "ideas" to spelling (Diederich, 1974). In contrast, analytic text analysis schemes such as T-unit analyses or Halliday and Hassan's measures of cohesion focus on isolated components of the written piece (Halliday & Hassan, 1976). Diederich's "flavor" subscale is far more qualitative and judgmental than counts of numbers and types of cohesive ties. In classroom evaluations of student writing, grades and teachers' comments, too, may reference a range of essay features such as content, organization, and mechanics (Freedman, 1979); or comments may only relate to sentence-level problems.

One issue in developing or using a rating scheme is the meaning of writing score(s). From a psychological perspective, does being a "2" vs. "4" discriminate between levels of a student's writing competence? At present, there is little research evidence that any sets of criteria in actual use are more valid than others for discriminating between levels of expertise. From a logical perspective, how specific, replicable, and informative are rating criteria? Pedagogically, what implications do the scores have for diagnosing strengths and weaknesses? The bases of the score, the criteria, should serve as feedback to teachers, students, and parents. To be fair, criteria employed in minimum competency tests should specify writing elements that are basic writing skills, e.g., organization, support, mechanics. The criteria should also be those amenable to instructional intervention. The more judgmental, qualitative, sophisticated, and less teachable writing elements such as flavor, style, or voice would seem less fair and less useful, and would therefore be inappropriate as rating criteria for judging basic writing competence. Specification of criteria may be the most important decision affecting the utility of information provided by assessment, both large scale and classroom level. Certainly, consensual decisions on these criteria should involve instructional and evaluation personnel.

It seems logical that criteria used in large scale writing competency assessment should reflect, if not derive from, criteria used to evaluate student classroom writing. An ideally integrated instructional system, one which targets particular writing elements as important basic competencies, would involve teachers and evaluators in specification of rating criteria and encourage focused classroom guidance, feedback, and evaluation on these elements. Instructionally, specification of valued basic criteria could provide a more comprehensive framework for teachers to focus instruction and communicate feedback to students about their writing. The scanty research on classroom evaluation methods suggests that teacher comments more often cite easily identified sentence-level mechanical errors than text level feedback such as organization and support (Pitts, 1978; Quellmalz, Baker, & Enright, 1980). As Coffman pointed out, while few would recommend complete restriction and regulation of the criteria teachers use in classroom writing assessment, neither would they condone subjecting students and the instructional program to wildly fluctuating, idiosyncratic standards of individual teachers (Coffman, 1971). Some standardization of writing criteria seems particularly critical for minimum competency goals. And, of course, schools using the same criteria for system-wide and classroom assessment would eventually reduce the cost of training raters, Assuming that logical, fair, and useful assessment criteria have been specified, the format for recording scores remains a problem. Many large scale assessments report a single, holistic score. A logical question is whether it makes sense to comment on component features of a student's writing instead of, or in addition to, its overall quality. A likely question to be raised about a single global score by a teacher, student, parent (or lawyer) is "Why?" followed by "Show me." While writing theory may suggest that the "whole" is greater than the sum of its parts, research in psychology and pedagogy suggests that learners advance when taught how to use components and combine them into competent performance (e.g., Skinner, 1957; Resnick, 1980). Another logical question is whether students are differently classified as masters and non-masters and/or if analytic schemes yield a differential score profile. Winters (1978) found that various scoring rubrics including a general impression scale, two analytic scales and a T-unit analysis, did classify students differently. Quellmalz, Smith, Winters & Baker (1980) found that three separate holistic rubrics and an analytic rubric classified entering freshman differently. Similarly, Polin (1980) found very low correlations between primary trait and analytic ratings of the same essays. Each of these studies compared scoring rubrics which referenced some similar criteria but which, in application, produced variable characterizations of the same essays. Still unexamined are the cost benefits of scales using the same criteria, but recording a single, holistic judgment vs. several separate analytic scores. Such a study is currently in progress (Quellmalz, 1981).

A major problem for large scale writing assessments, to be sure, is the cost of providing detailed ratings. In the narrowest sense, cost is measured in terms of time required to train raters and time required to rate papers. Generally, training on more criteria that are very explicit requires more time than training on fewer or less explicit criteria.

Currently available data on scoring costs indicate that training time for holistic and primary trait scoring averages two to four hours (Powliss, Bowers, & Conlan, 1979; Mullis, 1980), and for analytic scoring averages six to eight hours (Smith, 1978; Quellmalz & Capell, 1979). Trained raters can reliably assign a holistic or primary trait score to a student's paper in 30 seconds to 1½ minutes (Powliss et al., 1979; Mullis, 1980). Rating time for providing five to eight separate analytic scores range from four to five minutes for multi-paragraph essays and from two to four minutes for paragraphs (Smith, 1978; Quellmalz & Capell, 1979).

In a recent study comparing two score formats -- an analytic scheme or a holistic scheme modified to provide diagnostic checks for students rated below mastery -- Quellmalz found that average rating times per paper differed by approximately one minute (Quellmalz, 1981). Is the additional training and rating time "worth it?" School systems weighing this question might consider broader definitions and implications of cost. First, the cost of either analytic or holistic training could be jointly shared as an inservice activity by curriculum budgets. These training costs would also then decrease when all teachers in a system were trained and thus would require only periodic review of the procedures. A second potential cost sharing strategy is to view essay ratings as diagnostic components of the instructional system to both focus and monitor program improvement. A third cost concern is an ethical one. Students spend considerable time producing writing samples and the psychological and opportunity costs of making uninformed or erroneous decisions of student failure can be profound. Finally a system might consider the degree of specific support useful for defending mastery/non-mastery classifications; the costs of remediation and lawsuits because of misclassifica.ions can be high.

PROBLEM 4: TECHNICAL QUALITY OF RATING CRITERIA

A fundamental responsibility of an assessment program is the documentation of its technical quality. For writing assessments this becomes a problem of scale stability and validity, i.e., demonstrating that score criteria are applied uniformly within and between rating occasions and that other measures of student writing competence corroborate the test ratings (Quellmalz, 1980).

When carefully structured scale training sessions precede actual rating, most holistic and analytic rating scales can demonstrate high inter-rater reliability (Powliss et al., 1979; Mullis, 1980; Quellmalz, 1980; Steele, 1979; Van Nostrand, 1980). But inter-rater agreement within a rating session is not sufficient for demonstrating scale reliability. Analogous to the problem of test-retest reliability, a reliable scale must be stable, i.e., demonstrate that its criteria would be applied consistently by new sets of raters to both a new set of papers and to the set of papers scored by the first raters. To the extent that criteria are differently applied, the scale is not stable and reliable (Quellmalz, 1980).

Few scales currently used in writing assessment report data about their stability across sets of raters and rating occasions. It seems that scales with more explicit and operational criteria are less susceptible to fluctuating judgments and are more likely to be stable across paper sets and raters. Holistic scales such as the ETS method, which awards scores according to a paper's ranking within a unique set of papers, result in a sliding scale (Conlan, 1979). A "2" paper in one paper set may well have characteristics quite different from a "2" paper in a set of papers with a broader or narrower quality range. While some attempt is made to stabilize judgments across sets of raters by inserting anchor papers during training, anchor papers are less frequently interspersed in actual rating sequences. Statistical evidence of the comparability of scores given on any such anchor paper by different groups of raters is noticeably, and seriously, absent. Thus, holistic scales using ranking procedures within sets and unexplicated criteria are suitable for norm-referenced selection decisions, but can not meet competency test requirements for stable, uniform application of criteria. On the other hand, holistic scales based on more descriptive criteria, such as the primary trait method (Lloyd-Jones, 1977), may be more likely to permit stable application across paper and rater sets. Reports for most analytic scales also document inter-rater reliability within rating occasion but do not track stability across occasions. For analytic as well as holistic scales, precision of criteria is a critical factor in achieving scale stability. School systems designing writing assessments should routinely report inter-rater reliability and check scale stability on common paper sets scored at different rating sessions. These measures will reassure stakeholders that assessments are uniform and fair.

The task of documenting the validity of writing assessment rating scales can take several forms. Most competency-based writing assessments attempt to establish content validity through expert judgments about the skills assessed (Breland & Ragosa, 1976). Few writing assessment programs go on to subject the rating scales used to evaluate those skills to content validity scrutiny. Since, for written production, the scale defines what acceptable writing is, the content validity of scales should be judged by the same procedures as test items or specifications. It may be that some scales with vague criteria or criteria heavily weighted toward sentence-level mechanics would not get the stamp of approval from a broad range of experts. It should be noted that holistic scales with no explicit criteria are "content" free and assignment specific. These scales are not suitable for competency assessments.

Of course, content validity is only one index of validity (Cronbach, 1971; Messick, 1975). Concurrent or predictive and construct validity should also be examined. The most common method for validating large scale rating schemes has been to report their correlations with other writing-related measures including other English grades, reading test scores, and multiple choice writing test scores. Many of these "criterion" variables, however, are even more questionable indicators of writing ability than the rating scale being validated. A major problem in validating rating scales is identifying appropriate criterion groups and test scores (Winters, 1978; Quellmalz, Spooner-Smith, Winters, & Baker, 1980). A directly related criterion would be relationships of immediately preceding and subsequent writing assignment scores. Unfortunately, as different criteria are often employed in other rating scales and/or in teachers' grading of assignments, few appropriate direct comparisons are possible.

From the student's viewpoint, this problem raises concerns for fairness and instructional validity. How closely do the criteria used in the assessment match those used in the classroom, and how closely do they represent writing skills for which the student has received instruction? Fundamental precepts of fairness require that if a system hasn't explicitly taught the skills, it shouldn't hold the student accountable for being competent in these skills. For example, originality, humor, and flavor are desirable features of writing; they are not often directly taught. If we have no information on the criteria used in holistic scoring, that method isn't fair; we have no way to determine if what was tested was what was taught. The legal implications of this dilemma are obvious.

SUMMARY

Balancing ideally detailed analyses of students' writing with the costs of those analyses is no easy task. School systems and teachers across the country are wrestling with the problem and arriving at varying solutions. Some systems don't even try to initiate large scale rating of writing samples. Some teachers assign little writing and provide cursory or global feedback. Other systems are willing to pay the price and mount articulated writing assessment and instructional systems (e.g., Detroit, Los Angeles, Pittsburgh).

Some rating schemes apply explicit, replicable, reasonable criteria; some scales are silly, some are misapplied, some are downright harmful.

Large scale assessments can devise ways to reduce the costs of training raters to score large numbers of essays. In an ideally integrated assessment system, tasks and criteria for the large scale assessment would be the same as those used in the classroom. A district or state might construct a scale that referenced basic text components used by classroom teachers, e.g., main idea, coherence, support, mechanics, and devise a scoring system which checks off papers as competent on each skill and also checks off in more detail the components falling below mastery. For example, one paper might have competent support and receive a mastery check; another essay might not and get a check because "details are not related to the main point," or "details are not concrete."

Systems might allocate the cost of training raters to staff development. All teachers could be trained in applying the rating criteria which should promote greater articulation of the formal assessment with classroom criteria. Districts such as Detroit find it cost effective to pay lay personnel to rate writing samples. Alternately, the system might ask teachers to swap papers. Teachers could use the rating scale to score writing of other students in the district in return for having their students' writing scored by other teachers trained as raters. This would reduce training costs for district scoring. Many alternative logistics could be engineered to spread the time and energy costs efficiently within existing system resources.

Critics of writing assessment are questioning the fairness and utility of these assessments. Too many school systems cite cost as the reason that they cannot provide more valid, useful assessment. We think the technology and ingenuity exists to devise more defensible writing assessments now. We should no longer permit concern for cost to outweigh concern for fairness and utility.

REFERENCES

Baker, E. L., & Quellmalz, E. S. Issues in eliciting writing performance: Problems in alternative prompting strategies. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA, April 1980.

Bourne, L. J. Human conceptual behavior. Boston: Allyn & Bacon, 1966.

Breland, H. M., & Braucher, J. L. Measuring writing ability. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Breland, H., & Ragosa, D. Validating placement tests. Paper presented at the annual meeting of the American Educational Research Association San Francisco, 1976.

Britton, J. The composing process and the functions of writing. (Chapter 2) In Cooper and Odell (Eds.), Research on composing: Points of departure. Urbana, IL: National Council of Teachers of English, 1978.

Cazden, C. B. Two paradoxes in the acquisition of language structure and functions. In K. Connolly & J. S. Bruner (Eds.), The growth of competence. New York: Academic Press, 1974.

Coffman, W. E. Essay exams. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Conlan, G. Comparison of analytic and holistic scoring techniques. Princeton, NJ: Educational Testing Service, 1979.

Cooper, C. R. Current studies of writing achievement and writing competence. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.), Washington, D. C.: American Council on Education, 1971.

Crowhurst, M. Syntactic complexity in narration and argument at three grade levels. Canadian Journal of Education, 1980.

Diederich, P. B. Measuring growth in English. Urbana, IL: National Council of Teachers of English, 1974.

Florio, S. Learning to write in the classroom community: A case study. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Follman, J. C., & Anderson, J. A. An investigation of the reliability of five procedures for grading English themes. Research in Teaching of English, 1967, 190-200.

Freedman, S. How characteristics of student essays influence teachers' evaluation. Journal of Educational Psychology, 1979.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Halliday, M. A., & Hassan, R. Cohesion in English. London: Longman, 1976.

Lloyd-Jones, R. Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), Evaluating writing: Describing, measuring, judging. Urbana, IL: National Council of Teachers of English, 1977.

Messick, S. A. The standard problem: Meanings and values in measurement and evaluation. American Psychologist, 1975, pp. 955-966.

Moss, P., Cole, N., & Khampalikit, C. A comparison of direct and indirect writing assessment methods. Journal of Educational Measurement, in press.

Mullis, J. A. Using the primary trait system for evaluating writing. National Assessment of Education Progress, 1980.

National Assessment of Educational Progress. Reading, thinking and writing: Results from the 1979-80 National Assessment of Reading and Literature. Denver, Colorado, 1981.

Pitts, M. Relationship of classroom instructional characteristics and writing in the descriptive/narrative mode. Center for the Study of Evaluation, University of California, Los Angeles: CA, November, 1980.

Polin, L. Alternative conceptions of the writing skill domain: Problems for the practitioners. Paper presented at the National Council on Measurement in Education, Boston, 1980.

Powliss, J. A., Bowers, R., & Conlan, G. Holistic essay scoring: An application of the model for the evaluation of writing ability and the meaurement of growth in writing ability over time. Paper presented at the annual meeting of the American Educational Research ASsociation, San Francisco, 1979.

Praeter, D., & Padia, W. Effects of modes of discourse in writing performance in grades four and six. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Quellmalz, E. S. Domain-referenced specifications for writing proficiency. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1978.

Quellmalz, E. S. Assessing writing proficiency: Designing integrated multi-level information systems. Paper presented at the annual meeting of the National Reading Conference, San Diego, CA, 1980.

Quellmalz, E. S. Report on Conejo Valley's Fourth-Grade Writing Assessment: Fall, 1981.

Quellmalz, E. S., & Capell, F. Defining writing domains: Effects of discourse and response mode. Report to the National Institute of Education, November, 1979. (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation.)

Quellmalz, E., Spoc er-Smith, L. S., Winters, L., & Baker, E. L. Characterizations of student writing competence: An investigation of alternative scoring systems. Paper presented at NCME, April 1980. (Grant No. OB-NIE-G-79-0213 to the UCLA Center for the Study of Evaluation, 1980.)

Quellmalz, E., Baker, E. L., & Enright, G. Studies in test design: A comparison of modalities of writing prompts. Center for the Study of Evaluation, University of California, Los Angeles, CA, November, 1980.

Resnick, L. What do we mean by meaningful learning? Invited address at the annual meeting of the American Educational Research Association, Boston, 1980.

Skinner, B. F. Verbal behavior. New York: Appleton, 1957.

Scribner, S., & Cole, M. Unpackaging literacy. Social Science Information, 1978, 17, 19-40.

Smith, L. S. Investigation of writing assessment strategies. Report to the National Institute of Education. Center for the Study of Evaluation, University of California, Los Angeles, November 1978.

Steele, J. M. The assessment of writing proficiency via qualitative ratings of writing samples. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, 1979.

Van Nostrand, A. D. Writing instruction in the elementary grades: Deriving a model by collaborative research. Providence, RI: Center for Research in Writing, 1980.

Veal, L. R., & Tillman, M. Mode of discourse variation in the evaluation of children's writing. Research in the Teaching of English, 1971, 5, 37-45.

Winters, L. The effects of differing response criteria on the assessment of writing competence. Report to the National Institute of Education, November 1978. (Grant No. OB-NIE-G-78-0213 to the Center for the Study of Evaluation.)